**Why is this an important area of study? [0.5 points]**

Performance comes with trade-offs. So benchmarking not only let us measure performance quantitively, it also allows us to determine what trade-offs are acceptable.

**Describe two different techniques/approaches discussed [0.5 points]**

1. Reporting Expected Validation Performance (EVP, best evaluation metric number we can get with current budget) as with regard to the computation budget (training time, number of hyperparameters, …). This measures the trade-offs between training efficiency and inference performance.
2. Roofline model: It’s a visualization that shows the peak performance that we might expect from a model with a given arithmetic intensity. It measures the upper bound of memory or computation performance.

**Discuss relative strengths [0.5 points] and weaknesses [0.5 points] of the two techniques described above. [1 point total]**

Technique 1 presents which model has the highest performance potential given some budget. But reporting the top-1 test metric is not robust to noise.

Technique 2 can tell us if the model is fully utilizing all the hardware available. But this doesn’t reflect potential performance difference introduced by increasing/decreasing batch size or other factors.